The Econometrics Journal

Econometrics Journal (2021), volume **00**, pp. 1–31.
https://doi.org/10.1093/ectj/utab021

Large-scale sport events and COVID-19 infection effects: evidence from the German professional football 'experiment'

PHILIPP BREIDENBACH[†] AND TIMO MITZE[‡]

[†]*Research Data Center, RWI—Leibniz Institute for Economic Research, Hohenzollernstraße 1–3, Essen, 45128, Germany.*
Email: Philipp.Breidenbach@rwi-essen.de

[‡]*Department of Business and Economics, University of Southern Denmark, Campusvej 55, Odense, 5230, Denmark.*
Email: tmitze@sam.sdu.dk

First version received: 6 April 2021; final version accepted: 31 May 2021.

Summary: This paper studies the effects of large-scale sport events with live spectators on COVID-19 infection trends at the local population level. Specifically, we compare the development of incidence rates in 41 German Nomenclature of Territorial Units for Statistics level 3 (NUTS-3) districts hosting a professional football match with at least 1,000 spectators vis-à-vis similar districts without hosting a match. Our empirical analysis builds on difference-in-difference and dynamic event study estimation for panel data. Synthetic control method is applied as a robustness check. While our findings generally do not point to significant treatment effects for the full sample of match locations, we find some noteworthy exceptions. Districts hosting first league matches with spectator attendance above the median (> 6,300 persons) and, particularly, matches without strict face mask requirements experienced a significant relative rise in incidence rates 14 days after the match. We also find that intra-district mobility increases on match days in treated districts, highlighting the significance of professional football matches as mobility-based infection transmission channel.

Keywords: *COVID-19*, *large-scale sport events*, *football*, *live spectators*, *face masks*.

JEL codes: *C23*, *I18*.

1. INTRODUCTION

When the first wave of the novel Severe Acute Respiratory Syndrome Coronavirus 2 Disease (COVID-19) pandemic hit the world in spring 2020, policymakers all over the world reacted with strict nonpharmaceutical interventions to suppress the spread of the virus. One particular intervention was the cancellation or postponement of large-scale sport events with live spectators (such as the Olympic Games in Tokyo and the European Football Championship 2020) given the fear that these events may turn into super-spreading events (De Bruin et al., 2020; Parnell et al., 2020). With declining infection numbers during the summer of 2020, restrictions on professional sport events were temporarily lifted in most European countries—also allowing fans back into the stadiums. Although the experience from this reopening could provide important insights for future health policy decisions, yet, very little is known about the potential infection effects from these reopened events. This is what we address in this paper.

Whereas athletes, clubs, and organisers obviously have a strong interest in continuing their professional activities, among other aspects, as a means to avoid detrimental revenue losses (see, e.g., Drewes et al., 2021, for the financial implications of COVID-19 on professional football), the broader public debate on the reopening of professional sport events is also concerned with the proper health precautions for players, officials, live spectators, and, ultimately, the society in general. That the political decision about the reopening of sport events for live spectators is not an easy one is reflected in two very different political statements about the prospects of organising large-scale sport events during the summer of 2021. While the Danish government announced on March 25 that at least 11,000 fans shall be allowed to football matches staged in Copenhagen during the rescheduled European Football Championship in June 2021,[1] the Japanese government has decided to make only a 'last minute' decision on whether domestic visitors shall be allowed to watch the Olympic Games live or not. Even a full cancellation of the Olympic Games seems still to be possible two months ahead of the planned opening ceremony on July 23 in light of rising COVID-19 cases in Japan throughout May 2021.[2]

To support public health policy decisions in assessing the infection risks of large-scale sport events with live spectators, this paper studies COVID-19 epidemiological trends associated with the German professional football 'experiment' during the first phase of the 2020/2021 season in late summer and early autumn 2020.[3] This experimental phase permitted a varying number of up to 10,000 spectators to watch football (soccer) matches live in the stadiums following explicit and publicly communicated hygiene protocols. We believe that German professional football is a suitable case study for our endeavour, as the setup allows us to identify epidemiological trends in a precisely defined eight weeks event window throughout September and October 2020. It is important to note, though, that our analysis is concerned with the reopening phase after the first pandemic wave in 2020—a time period for which neither rapid testing for a large amount of persons nor a vaccination against COVID-19 were yet available.

The focus of our analysis is on tracking epidemiological trends at the local population level in a quasi-experimental setup. In our empirical identification strategy we compare the local COVID-19 infection development in German districts—at the Nomenclature of Territorial for Statistics level 3 (NUTS-3)—hosting a professional football match with at least 1,000 spectators to those from a suitable comparison group of districts that did not host such an event. Our strategy builds on a broad methodical basis for statistical inference to obtain robust treatment effects. Specifically, we start out by running a series of difference-in-difference (DiD) estimations to compare average changes in COVID-19 infection rates between treated and nontreated NUTS-3 districts over time. We pay particular attention to the issue of regional heterogeneity as professional football clubs (and therefore match locations) are much more likely to be situated in bigger and more densely populated districts, which generally experienced higher infection rates during our sample period.

[1] See, e.g., a news report by the *Washington Post* at https://www.washingtonpost.com/sports/soccer/fans-to-be-allowed-into-matches-in-copenhagen-at-euro-2020/2021/03/25/ccd51c6a-8d61-11eb-a33e-da28941cb9ac_story.html (last accessed: March 25, 2021).

[2] See, e.g., news reports by the Spanish *AS* news website at https://en.as.com/en/2021/05/18/other_sports/1621338645_150850.html (last accessed: May 19, 2021) and the *New York Times* at https://www.nytimes.com/2021/05/25/sports/olympics/tokyo-olympics-cancel.html (last accessed: May 27, 2021).

[3] Politicians and representatives from the German Football Association (DFL) have repeatedly called this a 'testing phase' and 'experimental phase' for large-scale sport events, which is why we refer to the 'German football experiment'; see also https://www.n-tv.de/sport/fussball/Pandemie-Testphase-im-Stadion-So-tritt-die-Bundesliga-gegen-Corona-an-article22044930.html (last accessed: November 19, 2020). However, we want to stress that the experimental phase did not involve any randomised trial elements associated with the reopening to systematically assess drivers of the infection dynamics. We do so here by means of quasi-experimental tools.

We will account for this and other sources of regional differences not related to the treatment in focus, i.e., hosting professional football matches.

Besides a general juxtaposition of average COVID-19 incidence rate differences in treated and comparison regions, we are also interested in investigating the dynamic, i.e., time heterogeneous, nature of these epidemiological trends. As an extension to the baseline DiD estimates, we therefore implement a dynamic panel event study (PES) that allows us to study how daily treatment effects evolve over time. As previous COVID-related research has shown, considering dynamic effects is an important means to account for a latent incubation time, on the one hand, and accelerating infection dynamics over time, on the other hand. Both aspects may render average static effect estimates less conclusive. DiD and PES regressions are carried out in a two-way fixed effects framework and account for the staggered treatment start across match locations as well as latent confounding factors.

We apply synthetic control method (SCM) estimation as a robustness check. The SCM explicitly matches treated and nontreated districts based on their pretreatment trends, including mobility patterns and prior COVID-19 incidence rates, and is thus less sensitive to the common trend assumption underlying DiD and PES estimation. By combining all three methods for the empirical estimation of infection effects of professional football matches with spectators, we argue that this ensures the robustness of our identification approach. To our knowledge, this is the first paper that analyses the potential infection effects of large-scale sport events in times of COVID-19 at the local population level.[4]

Foreshadowing key results, for our full sample covering all matches with at least 1,000 spectators in all three professional football leagues together with the first round of the German national cup (DFB Pokal), we generally do not find persistent evidence for significantly higher COVID-19 infection dynamics in the treatment group vis-à-vis comparisons regions after treatment start. Although our estimates point to somewhat higher infection rates in treated regions, in almost all cases this difference is not statistically significantly different from zero for our full sample. Regarding these nonsignificant results, one concern is obviously that our district-level identification approach may suffer from a low power of detecting effects from individual football matches at the local population level. To address this concern, we have additionally estimated relative mobility effects on matchdays in NUTS-3 districts with matches showing that these regions have increased mobility rates vis-à-vis those without matches. We take this as an indication that professional football matches with at least 1,000 spectators are a significant event in treated NUTS-3 districts and that it is reasonable to argue that potential infection effects should become visible at the local population level (if present).

Different from the overall results, we find evidence for a statistically significant and dynamically evolving increase in the population-level COVID-19 incidence rate for first league matches. Analysing professional first league matches as a subsample of all matches is important as clubs in this league have by far the largest fan base, which may thus lead to higher mobility rates at match days and to stronger infection effects. These effects are particularly sizeable for first league matches with the number of spectators above median ($>$ 6,300 persons) and for matches in which the use of face masks was only required when entering/exiting the stadium but not throughout the entire match (i.e., when seated). Regarding the temporal evolution of effects, we find that

[4] An assessment of the health effects from the restart of German professional football for players from the two highest German leagues and officials working closely with them is reported in Meyer et al. (2021). The authors find that the implemented hygiene protocol involving regular PCR testing proved effective in avoiding COVID-19 transmissions. The conclusion for their study is that professional football training and matches can be carried out safely during the COVID-19 pandemic—though requiring strict hygiene protocols including regular PCR testing.

they turn significant approximately 1–2 weeks after the match took place, which is consistent with estimated incubation times and a reporting lag for the German COVID-19 data. Placebo treatment regressions for home districts of the respective visiting (away) teams of a match do not show any significant effect, which supports the causal interpretation of our results.

All in all, if we weigh the different estimation results against each other, we argue that certain large-scale sport events, i.e., those with a large fan base that accordingly mobilise many people to watch these events live, may pose an additional COVID-19 infection risk under the circumstances studied. Such effects are most significant for estimators that account for the time dynamics in incidence rate developments at the local population level. Importantly, our results also point to policy implications in light of these findings. They suggest that proper hygiene protocols are a key element to reduce the risk of higher COVID-19 infection rates for stadium visitors and the local population. Most prominently, they point to the importance of enforcing strict and permanent face mask regulations during stadium visits together with social distancing rules in the organisation of such events. Public health rules (along with rapid testing of stadium visitors, which was not available by the time of our analysis) should become even more important if further evidence underlines the higher transmissibility of novel variants of concern (VOC) of the Coronavirus 2 (see, e.g., Davies et al., 2021) and if vaccination strategies have not yet established herd immunity.

2. INSTITUTIONAL BACKGROUND, DATA, AND VARIABLES

2.1. Institutional background

The first professional football matches with live spectators after the national lockdown in spring 2020 took place throughout September 2020 and covered the German national cup and the three professional football leagues. During that time period, the spread of COVID-19 was largely under control and daily infections rates were low to moderate. For instance, the seven-day incidence rate as key epidemiological indicator in Germany varied between 10 and 15 cases per 100,000 inhabitants (over the last seven days) throughout this period. Nonpharmaceutical interventions were less restrictive compared to the lockdown phase in spring but were still in place. Specifically, in most federal states, only up to ten persons were allowed to meet in a public space, with some exemptions being granted for amateur sport activities. For example, in North Rhine-Westphalia (the largest German state with the most professional football clubs), shops, restaurants, bars, and other private and public establishments were open, but with a maximum number of customers and with strict face mask obligations for any indoor activity in public (CoronaSchVO, 2020).

These prevailing public health regulations illustrate the unique role of professional football matches with many spectators in terms of gathering of people. In response to a growing public demand for these activities, the federal government and state-level governments agreed on an experimental test phase over six weeks, which allowed stadium visits by fans under the following conditions. First, the seven-day incidence in districts hosting a match should be below 35 per 100,000 inhabitants in the days before the match.[5] Second, the stadium capacity could only be utilised up to a certain limit, typically 20−25% of the maximum stadium capacity. Third, in order to keep inter-regional mobility low, fans from away teams were generally not allowed to enter the stadiums.

[5] This implied that some professional football matches during this test phase took place as ‘ghost games’ without live spectators, see Table S3 in the Online Appendix for details on first league matches.

In addition, organising clubs were requested to immediately apply changes to public health regulations issued by German federal states and, in general, local authorities requested individual hygiene protocols taking critical contact points such as arrival in the stadium, ticket personalisation, and the spatial distribution of spectators within the stadium into account. These local hygiene protocols also included possible face mask obligations for attending a match. While these hygiene rules should generally avoid a close crowding of fans inside the stadium, they are not always effective to ensure distancing in reality.[6] Moreover, while most rules were common for all match locations, importantly, certain variations in the design of the specific hygiene protocols for individual matches could be observed due to regulations by local authorities. This enables us to test for different channels of potential disease transmission.

2.2. Data and variables

Our analysis uses district-level data (NUTS-3 level) on the number of newly reported COVID-19 cases provided by the Robert Koch Institute (RKI), which is in charge of infectious disease surveillance in Germany. The RKI publishes the most recognised COVID-19 database for Germany based on information fed into the database by local health authorities (RKI, 2021). Data can be accessed freely via the RKI dashboard at: https://npgeo-corona-npgeo-de.hub.arcgis.com. The database offers, among others, information on new infections for each NUTS-3 district on any given day.

While our district-level identification approach thus differs from micro-methods that directly try to retrace infection chains of persons tested positive for SARS-CoV2 on the basis of tracking and outbreak reports, we believe that our regional focus is a suitable choice for two reasons.[7] First, in the German public health monitoring system, individual transmission chains can be traced back to specific events only for a small (and potentially systematically biased) fraction of reported cases (RKI, 2020a, 2020b). This leaves room for latent transmission channels. Accordingly, statements from local health authorities, such as in Berlin, which report to have found no evidence for reported infections from stadium visitors to Hertha BSC and 1.FC Union Berlin based on such tracking reports, should be interpreted carefully.[8] A district-level analysis implicitly captures all such latent transmissions that take place at the local level. Second, the district level also gives us the opportunity to detect potential infection chains which are only indirectly linked to the stadium visit, such as the intensity of public transport use, gatherings outside the stadium, and mobility during match days in general. This may provide important insights for public health policy.

We merge the COVID-19 data with district-level, cross-sectional information (population, population density, age structure, local health system, etc.) obtained from the INKAR online database of the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR, 2021). We also collect data on daily temperatures per NUTS-3 district from Deutscher Wetterdienst (DWD, 2021) and data on daily mobility changes per NUTS-3 district from the German Statistical Office (Destatis, 2021). Mobility changes (in percent) based on individual mobile phone data are computed as the difference in mobility patterns between a specific calendar date and the average monthly value for the corresponding weekday in the previous year. For

[6] See, e.g., a press report from *Westfalenpost* showing celebrating fans of Bayern Munich (without social distancing) after winning the Supercup: https://www.wp.de/sport/bayern-sieg-mit-nebenwirkung-fans-sorgen-fuer-corona-aerger-id230520196.html (last accessed: November 14, 2020).

[7] See Tupper et al. (2020) for an event-based identification approach from such outbreak reports.

[8] See a report by the German-wide sports news portal *Kicker* at: https://www.kicker.de/behoerden-keine-corona-infektionen-bei-hertha-und-union-heimspielen-nachweisbar-788386/artikel (last accessed: November 14, 2020).

instance, a value of −20 shows that mobility for a given day in 2020 was 20% lower than the average mobility for the same weekdays in the respective month of 2019.[9]

Data on the timing of professional football matches together with information on the number of spectators are obtained from the German Football Association (Deutscher Fussball-Bund; see DFB, 2021a) website. Data on stadium size, car parking capacities, specific hygiene protocols (e.g., with regard to the maximum number of spectators permitted; social distancing measures applied, such as mask wearing protocols and sequence of admittance of fans into the stadium, etc.) are collected from the websites of the individual football clubs and further online resources accessible through the DFB data centre (DFB, 2021b). All study data used for our analysis are publicly available and are provided as supplementary information to facilitate replication studies.

An outcome variable of interest is the rate of new infections by NUTS-3 district and day defined as

$$IR_{i,t} = \frac{Cases_{i,t}}{Population_i} \times 100{,}000,$$

where $Cases_{i,t}$ is the number of newly reported COVID-19 cases in region i at time t and $Population_i$ is the district's population level (measured as of December 31, 2019); $IR_{i,t}$ is typically referred to as (daily) incidence rate. While our analysis mainly uses this incidence rate as key outcome variable, we also define the seven-day incidence rate as the sum of daily incidence rates over the last seven days per 100,000 of local population. The seven-day incidence is used in Germany specifically for disease surveillance and for public policy decisions. In several instances, we also use the seven-day incidence to compute descriptive statistics and robustness tests. The advantage of the seven-day incidence rate as a rolling indicator is that it is less sensitive to outliers; however, at the same time it is less precise in identifying daily effects associated with a certain event. Our choice for using one of the two incidence rates for specific empirical tests is closely linked to our empirical identification strategy and the choice of estimators.

3. ESTIMATION STRATEGY

3.1. Overview

Our estimation strategy is based on three pillars. We start with a baseline DiD approach (first pillar) that compares the average development in COVID-19 incidence rates over time (pre- vs. post-match) and across regions (districts with vs. districts without matches). These static DiD estimates should be seen as a first and straightforward baseline test for the significance of potential infection effects of professional football matches at the local population level. Subsequently, we extend this approach in two directions: (a) we run a dynamic PES (second pillar), which allows us to place a stronger focus on (i) the identification of dynamic effects and (ii) an intra-group comparison between treated districts hosting a match at different points in time. (b) We apply the SCM (third pillar), which places a particular emphasis on matching treated and nontreated districts based on their pretreatment COVID-19 trajectories and further structural characteristics. It is thus less sensitive to the common trend assumption in DiD and PES estimation. Estimator details are given below.

[9] See also Schlosser et al. (2020) for an analysis of COVID-related regional mobility changes in Germany.

3.2. Difference-in-difference estimation

3.2.1. General setup. We chose DiD estimation as our baseline specification as it is a well-established tool for estimating average treatment effects on the treated (ATT) (e.g., Lechner, 2010). A recent discussion of the use of DiD estimation to identify causal effects of COVID-19 policies is given in Goodman-Bacon and Marcus (2020). Here we apply the DiD estimations within the general class of two-way fixed effects (FE) model specifications, which allows us to account for latent structural differences between regions (time constant) and further observed common time varying confounding factors other than the treatment event in focus (see, for instance, Cho, 2020, for a similar approach assessing lockdown effectiveness related to COVID-19 at the country level).[10]

It is important to point out that the DiD approach taken here already accounts for the staggered treatment start across cross-sectional units (see, e.g., Athey and Imbens, 2018; Goodman-Bacon, 2018; Borusyak and Jaravel, 2020). While some scholars, such as Sun and Abraham (2020), distinguish an event study from the DiD approach based on the presence of a staggered adoption design, we here refer to a panel event study as a tool to estimate dynamic, i.e., time heterogeneous effects, rather than static treatment effects.[11] The DiD model proposed as our baseline specification accordingly corresponds to a static panel event study in the definition of Sun and Abraham (2020). This baseline specification with staggered treatment adoption can be written as

$$IR_{i,t} = \delta FTB_{i,t} + X_{i,t}\beta + \tau_t + \mu_i + \varepsilon_{i,t}, \tag{3.1}$$

where $FTB_{i,t}$ is the treatment indicator for football matches (see details below) and $X_{i,t}$ captures additional time varying control variables, in particular (linear and quadratic), region type-specific time trends, τ_t are common time fixed effects for sample days, and μ_i controls for time constant, region fixed effects, while $\varepsilon_{i,t}$ is the model's error term. Our focus is on estimating the coefficient δ, which essentially captures the impact of large-scale sport events on the (daily) incidence rate in the time period after the football match took place. A positive and statistically significant estimate would imply that hosting professional football matches with spectators increases, on average, the infection rate in treated vis-à-vis comparison regions. Equation (3.1) describes a static panel model setup with regard to the evolution of the outcome variable. As a robustness test, we also estimate (3.1) as a dynamic panel data model by including a one-period lagged value of the incidence rate among the regressors. This inclusion shall additionally control for the autoregressive nature of $IR_{i,t}$ apart from the above described time trends.

Our treatment indicator $FTB_{i,t}$ is a binary dummy that takes a value of 1 from the day onwards when the individual NUTS-3 district i hosts the first professional football match ($Match_i$) with at least 1,000 spectators, i.e., $FTB_{i,t} = 1[t \geq Match_i]$. Hence, we focus on an absorbing treatment such that the treatment status is a nondecreasing series of zeros and ones (Sun and Abraham, 2020). We incorporate all matches in our event-based definition of $FTB_{i,t}$ that took place from September 11 to September 28, 2020. These matches include the first round of the national cup (DFB-Pokal) and the first two match days of the first three professional football leagues (erste, zweite und dritte Bundesliga), respectively.[12] Limiting our treatment period to the first two match

[10] Such structural differences may occur from different testing strategies at the regional level and changes in nationwide testing procedures. Region and time fixed effects control for such differences.

[11] Sun and Abraham (2020) define a difference-in-difference design as a setting where units either receive their first treatment at a common time period t_0 or are never treated.

[12] Districts that have hosted more than one match during the sample period (i.e., those with more than one football club in German professional leagues) are defined as treated regions at the first match day.

days ensures that we do not measure multiple treatments per district because the clubs hosting a match at the first match day are playing in another city on the second match day (and would only host a new match the week after). Moreover, as default specification, all of our analyses are trimmed to a maximum time window of 14 days after treatment start as this date typically marks the beginning of the next (home) match day for each club. This post-treatment period is also applied in the subsequent PES and SCM estimations.

As it has been shown in detail in Mitze et al. (2020), a time lag of 14 days is generally sufficient to cover approximately 75% of reported infections associated with a specific day t in the RKI data given that incubation times and a reporting lag need to be considered. Hence, the chosen time windows for the treatment period should be sufficient to identify potential infection effects associated with the treatment. Robustness tests include a maximum lag of up to 20 days after the match day. A summary table including information on match days, matches, and spectators is given in Table S1 in the Online Appendix. Additionally, Figure S1 in the Online Appendix provides an overview of our sample organisation. As the table shows, our overall sample covers the period between August 10 and October 18, 2020.

3.2.2. Regional Heterogeneity. A closer inspection of the data indicates the necessity of an identification strategy that accounts for underlying differences between districts in the treatment and comparison group. This is shown in Table 1, which highlights basic characteristics of NUTS-3 districts together with t-tests for mean differences between groups. As the table reveals, treated districts, i.e., those with professional football clubs (across all three leagues) hosting a match with at least 1,000 spectators, have, on average, larger population levels and are also characterised by a higher population density than nontreated comparison regions indicating that treated regions are mainly urbanised centres. This difference also become evident if we compare the share of regions belonging to different structural region types as measured by the BBSR (2021). While all treated regions belong either to the group of large district-free cities (*kreisfreie Städte*; region type I) or urbanised districts (*Landkreise*; region type II), this is only the case for roughly 45% of comparison regions. None of the matches have taken place in districts (*Landkreise*) of type 3 or 4, which comprise (sparsely populated) rural areas in the definition of BBSR (2021). For these variables and further indicators of the regional demographic composition, the results of the reported t-tests point to significant mean differences across groups.

In addition to the higher mean incidence rate in treated NUTS-3 districts, a closer look at epidemiological trends for the four different region types illustrates that an approach which only covers time constant differences (i.e., district fixed effects) between groups may fail to provide credible evidence. Specifically, it may lead to a rejection of the common trend assumption underlying DiD estimation, which states that treated and nontreated districts would have followed parallel trajectories over time if treatment had not occurred (Lechner, 2010). Figure 1 shows the heterogeneity of the 7-day incidence rate (per 100,000 inhabitants) across the region types throughout September/October 2020, pointing to a stronger increase of the infections (in absolute terms) in NUTS-3 districts of type 1 and 2 in contrast to type 3 and 4. Given that all matches with spectators have taken place in regions of type 1 or 2, the figure underlines the need to account for this growing structural regional difference in our DiD estimations.

We propose three alternative estimation methods to overcome this problem. As a first solution, we reduce the number of cross-sections (regions) in the sample to cover only NUTS-3 districts of type 1 and 2. Following this idea, those districts (defining the comparison group) within this subsample in use are much more likely to be similar to the group of treated districts. However, the sample size decreases significantly following this approach. As a second

Table 1. Structural characteristics of treated and nontreated (comparison) NUTS-3 districts.

Groups	Treated mean (SD)	Comparison mean (SD)	Difference
Number of included NUTS-3 districts	41	360	
Incidence rate (new infections per 100,000 of population)	3.24	2.40	0.841***
	(0.078)	(0.028)	
Seven-day incidence rate (per 100,000 of population)	20.56	15.05	5.503***
	(1.578)	(0.491)	
Mobility changes (in %, relative to previous year)	− 0.169	− 0.033	− 0.137***
	(0.009)	(0.006)	
Average temperature (in degree Celsius)	13.74	13.04	0.707***
	(0.122)	(0.047)	
Population (in persons)	478,930.20	175,433.90	303,496***
	(13,408.820)	(945.871)	
Population density (persons per km^2)	1,703.86	400.49	1,303.4***
	(18.532)	(4.153)	
Region type 1 (large cities, *kreisfreie Städte*)	0.85	0.09	0.765***
	(0.008)	(0.002)	
Region type 2 (Urbanised districts, *Landkreise*)	0.15	0.35	− 0.201***
	(0.008)	(0.004)	
Region type 3 (rural districts, *Landkreise*)	0	0.28	− 0.281***
		(0.003)	
Region type 4 (sparsely populated districts, *Landkreise*)	0	0.28	− 0.283***
		(0.003)	
Share of females in population (in %)	50.81	50.57	0.236**
	(0.017)	(0.005)	
Average age females (in years)	44.22	46.059	− 1.840***
	(1.831)	(2.060)	
Average age males (in years)	41.35	43.381	− 2.029***
	(1.435)	(1.758)	
Old-age dependency rate (in %)	30.83	34.75	− 3.924***
	(5.408)	(5.328)	
Young-age dependency rate (in %)	19.924	20.609	− 0.684***
	(1.396)	(1.429)	
Physicians (per 10,000 of population)	18.44	14.148	4.294***
	(3.259)	(4.298)	
Pharmacies (per 100,000 of population)	27.815	26.911	0.904
	(3.976)	(4.978)	
Share of highly educated in population (in %)	22.21	12.02	10.195***
	(0.164)	(0.039)	

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$ are significance levels for *t*-tests on mean difference (null hypothesis: equal means); SD = standard deviation. Data on newly reported COVID-19 cases used to calculate incidence rates are taken from RKI (2021); mobility data are taken from the German Statistical Office (Destatis, 2021) and average temperature from the Deutscher Wetterdienst (DWD, 2021). All other regional variables are obtained from the INKAR database (Indikatoren und Karten zur Raum- und Stadtentwicklung) of the German Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR, 2021; data are extracted for the latest available sample year, 2017). See main text for further variable descriptions.

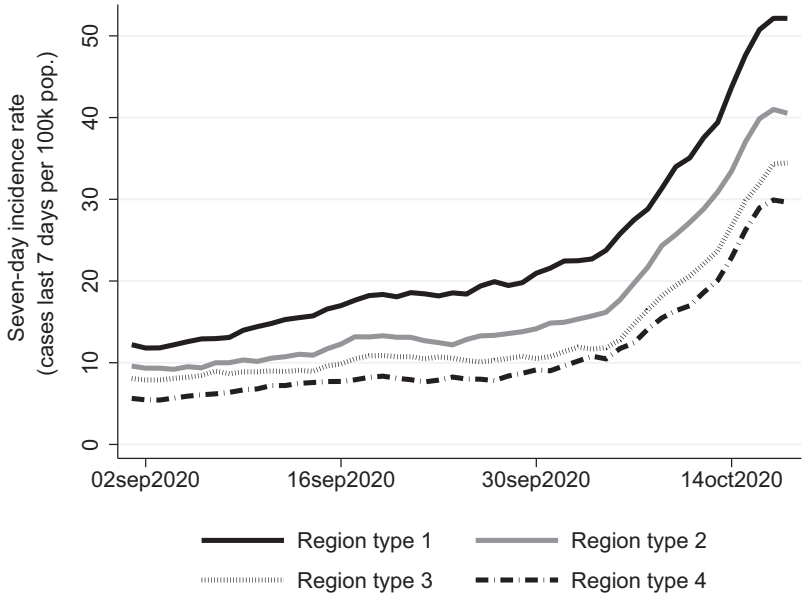

Figure 1. Temporal evolution of seven-day incidence rates by region types.
Notes: See Table 1 for further information on the categorisation of the four different region types.
Source: Own figure based on data from the INKAR database (Indikatoren und Karten zur Raum- und Stadtentwicklung) of the German Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR, 2021).

solution, we rely on the full sample and include region type-specific (linear and quadratic) time trends in addition to common day fixed effects (τ_t). These trends are able to capture different infection dynamics across region types while still allowing us to conduct inference based on the full sample of German NUTS-3 regions. However, trends which are (at least partly) based on the deviating development of the treatment group are prone to confound with treatment effects.

As a third solution, we apply a doubly robust (DR) estimator, which combines two alternative approaches of controlling for confounding factors in order to correctly estimate the treatment effect on the outcome (Funk et al., 2011). Specifically, we extend the standard regression-adjusted DiD approach as in (3.1) by inverse probability weighted (IPW) estimation (see, e.g., Heckman et al., 1998; Abadie, 2005; Sant'Anna and Zhao, 2020). The main idea of controlling for observed confounding factors in the IPW approach is to assign larger sample weights to nontreated districts, which have similar characteristics as treated districts during a time period before treatment starts. This, in turn, could make the common trend assumption more credible. One advantage of the DR-DiD approach is, hence, that it delivers consistent estimates if at least one of the two estimation approaches is correctly specified (Sant'Anna and Zhao, 2020). With respect to the analysis of treatment effects associated with COVID-related policy interventions, Goodman-Bacon and Marcus (2020) point to the particular importance of using reweighting schemes for sample units to balance pretreatment infection levels and trends across groups. Here we follow the main strand of the empirical literature on DR estimation that suggests a propensity score (PS) based parametric

reweighting approach. PS values are thereby obtained from a first-step probit model that specifies the likelihood of a district to be included in the treatment group, i.e., to host a professional football match with live spectators.

Regional sociodemographic characteristics, as shown in Table 2, as well as lagged values of the (cumulative) incidence rate are included in the first-step probit model, which is estimated as pooled specification for the pretreatment period from August 10 to September 10, 2020 (the first professional football match covered in our sample takes place on September 11, 2020). Detailed regression outputs for the probit model are given in the Online Appendix (Table S2) as well as tests for covariate balancing before and after PS-based reweighting. Having obtained a set of PS values with satisfactory balancing properties, we then reestimate (3.1) by means of weighted least squares (WLS) regression (e.g., Freedman and Berk, 2008). To account for uncertainties associated with PS estimation, as a robustness test to standard inference, we also employ a two-step bootstrap method to calculate treatment effects and associated standard errors. The advantage of this two-step procedure is that it includes PS estimation and sample weight specification in each bootstrap iteration (see Wooldridge, 2010, and Bodory et al., 2020, for a general assessment of bootstrap methods for weighting estimators).

Taken together, we argue that the chosen baseline DiD specification with staggered treatment start is flexible enough to allow us to check the robustness of the estimated treatment effects for alternative estimation specifications (e.g., dynamic panel estimation, doubly robust estimation) and varying sample settings. However, as outlined in Borusyak and Jaravel (2020), one potential problem of the baseline estimates is that effects may be biased if short- and long-run effects differ, i.e., if treatment effects have strong dynamics. In this setting, the authors suggest running unrestricted regressions which do not impose any restrictions on the dynamics of treatment effects post-treatment.

3.3. Panel event study

We do so by applying a dynamic panel event study (PES). Whereas our static baseline specification estimates a single time constant treatment effect, this may be problematic in dynamic settings in which the treatment only gradually impacts the outcome variable over time put then has a potentially accelerating effect. This may be a relevant effect trajectory, particularly for epidemiological data. Additionally, in switching to a PES design, we can overcome major problems of the definition of suitable comparison groups (shown above) where treated and nontreated regions are prone to develop differently over time. As shown in (3.2), we establish an estimation strategy in our panel event study that controls for different developments between both groups over the whole observation period without affecting the estimated treatment effect.

Dynamic PES setups have previously been applied in a broad variety of settings to estimate COVID-related dynamic treatment effects, such as school reopening in Germany (Isphording et al., 2020; Von Bismarck-Osten et al., 2020), university students travelling during the US spring break (Mangrum and Niekamp, 2020), and mass protests from the Black Lives Matter movement (Dave et al., 2020). The dynamic PES design, as shown below, seeks to identify daily treatment effects in the following manner

$$IR_{i,t} = \sum_{j=-N}^{M} \delta_j FTB_{i,t}^{j} + X_{i,t}\beta + \tau_{t,Treat} + \mu_i + \varepsilon_{i,t}, \tag{3.2}$$

where $\delta_j FTB^j_{i,t}$ is the j^{th} element of a set of binary, absorbing treatment indicators defined as

$$FTB^j_{i,t} = \begin{cases} 1[t \leq Match_i + j] \text{ if } j = -N \\ 1[t = Match_i + j] \text{ if } -N < j < M, \\ 1[t \geq Match_i + j] \text{ if } j = M \end{cases}$$

and the index $j = -N, \ldots, M$ denotes the maximum number of leads $(-N)$ and lags (M) considered for estimation (Schmidheiny and Siegloch, 2019). The inclusion of $-N$ leads explicitly tests for earlier effects in the outcome variable prior to treatment start. If such effects are significant and positive, a rise in the incidence rate in district i is likely driven by other latent factors, rather than by the hosting of a professional football match with live spectators. However, if those effects are absent before treatment start and coefficients for $FTB^j_{i,t}$ turn significant after this start for some of the included M lags, in our case, we take this as statistical evidence for significant infection effects of hosting a match. Plotting daily treatment effects δ_j allows us to identify the phasing-in on effects.

Similar to the static DiD baseline approach, we trim the panel to be balanced in relative periods, both 14 days before and after each football match, to arrive at a balanced panel in relative time. As Table 1 has shown significant differences between treated districts (hosting football matches) and nontreated districts, we model these differences in the estimation of treatment effect on incidence rates. Different from the baseline DiD specification, the PES allows us to control for heterogeneous daily developments in incidence rates between treated and nontreated districts. This is implemented by daily time effects for treated regions ($\tau_{t,Treat}$) in addition to the daily effects for all (treated and nontreated) regions (indicated by τ_t). The effects of the matches are nevertheless still identified as they are not measured via daily effects in the calendar time but via the daily effects in the relative time (depending on the temporal distance to the match) and we can fully control for different developments (even after the treatment start) between treated and nontreated districts. Implicitly, the treatment effects $FTB^j_{i,t}$ then indicate the differences at day i between treated regions which already hosted a match j days before and those which did not host a match in that temporal distance.

In the estimation of (3.2), the treatment indicator $FTB^j_{i,t}$ for the last pretreatment observation $(j = -1)$ is omitted to capture the baseline difference between treated and nontreated districts. Taken together, the panel event study approach as in (3.2) can hence be seen as an important extension and robustness check to our static baseline DiD estimation approach. However, both the static DiD and the dynamic PES approach chiefly depend on the validity of the common trend assumption. While we can cautiously use the daily PES results as a test for the presence of trend differences and anticipation effects before the treatment start, Sun and Abraham (2020) have pointed out that pre-trend estimates may be contaminated, i.e., that there is a nonzero correlation between the individual coefficients δ_j, which may result in a low power of this test.

3.4. Synthetic control method

We apply the synthetic control method (SCM) to check the robustness of the dynamic PES results. Similar to the PES approach, SCM can be used to track the evolution of treatment effects over time, where—in line with DiD and PES estimation—effects are estimated as changes in outcomes between a pretreatment and treatment period for treated and nontreated units. A key difference is, however, that the SCM approach does not rely on the common trend assumption as the presence of common trends between treated and nontreated comparison units is in itself a favourable factor for finding an appropriate counterfactual trajectory. The key identification approach of SCM is

thus to establish a counterfactual that mimics a situation in which the treatment, i.e., professional football matches, would not have taken place in treated districts. This is implemented by means of creating a synthetic control group consisting of the comparison districts chosen from a donor pool and by comparing the outcomes of treated units and the synthetic control after the start of the treatment.

The match between treated districts and the synthetic control group prior to treatment start is done through a minimum distance approach for a set of predictor variables evaluated along their pretreatment values for treated districts and all regions in the donor pool. This ensures that pretreatment differences in trends of the outcome variable are levelled. A formal description of the SCM approach and its underlying conceptual requirements is given in Abadie and Gardeazabal (2003); Abadie (2005); Abadie et al. (2010); Abadie (2020); Cavallo et al. (2013) among others. The Cavallo et al. study also extends single treatment SCM estimation to the case of multiple treated units as it is applied here. SCM has previously been applied to COVID-related research, for instance, to study the effect of face masks on SARS-CoV-2 infection numbers in Germany (Mitze et al., 2020), epidemiological trends associated with the emergence of SARS-CoV-2 variants of concern (Mitze and Rode, 2021) as well as lockdown effectiveness for a counterfactual of Sweden (Cho, 2020) and the United States (Friedson et al., 2021).

Here we essentially follow the approach presented in Mitze et al. (2020), which also builds on data for German NUTS-3 districts. We use a broad set of time varying and time constant predictor variables in the pretreatment period to construct the synthetic control group. Time-varying predictors include daily values for the incidence rate, the average temperature, and mobility changes in the last seven days before treatment start. Time constant predictors as chosen in Mitze et al. (2020) further comprise population density (population/square kilometre), regional settlement structures (categorial dummy), the share of highly educated in the population (in %), the share of females in the population (in %), the average age of females and males in the population (in years), old- and young-age dependency ratios (in %), the number of physicians per 10,000 of the population, and pharmacies per 100,000 of the population.

We apply SCM for multiple treated units and calculate a synthetic control group for each of the 41 NUTS-3 regions hosting a professional football match. The donor pool of controls is limited to nontreated districts out of the group of large district-free cities (*kreisfreie Städte*; region type I) or urbanised districts (region type II). As outlined above, this will ensure that predictor values of treated districts are not extremely relative to those values of comparison districts in the donor, i.e., that treated districts lie in the convex hull of comparison districts (Abadie, 2020).

Statistical significance of the estimated treatment effect is based on permutation. Specifically, we calculate 95% confidence intervals (CIs) from pseudo *p*-values obtained on the basis of comprehensive placebo-in-space tests. These tests calculate pseudo-treatment effects for all districts in the donor pool treating each of these districts as if it would have received the treatment. If the distribution of placebo effects yields only very few effects as large as the main estimate for treated units, then it is likely that the estimated effect is not observed by chance. One advantage of this exact permutation-based test is that it does not impose any distributional assumption on the model's errors (Abadie, 2020).[13] For multiple treated units, placebo effects are computed in such a way that each nontreated comparison unit is thought of as entering treatment at the same

[13] However, to be formally valid, this would ideally require a random assignment of treatment among units, a condition which is hardly met in observational studies. A solution to this problem would be to employ a permutation scheme that incorporates information in the data on the assignment probabilities for different units in the sample (Abadie, 2020). While it is beyond the scope of this paper, we hope that future work can further draw on this aspect.

time as the treated unit. Hence, two treated units with the same treatment start will have the same placebo sets.

Moreover, to account for differences in pretreatment match quality of the pseudo-treatment effects, only donor regions with a good fit in the pretreatment period are considered for inference. Specifically, we do not include placebo effects in the pool for inference if the match quality of the control region, measured in terms of the pretreatment root mean squared prediction error (RMSPE), is greater than 10 times the match quality of the treated unit (Cavallo et al., 2013). Based on the obtained pseudo *p*-values we calculate confidence intervals as described in Altman and Bland (2011). As the number of placebo averages becomes very large for multiple treated units, we conduct inference on the basis of a randomly drawn sample of 1,000,000 placebo averages (Cavallo et al., 2013).[14]

3.5. Estimation power

As described above, one concern for our identification strategy is related to the estimation power of our approach given that live spectators in stadiums only account for a fraction of the local population in a NUTS-3 district. Accordingly, analysing population-level epidemiological trends may not detect infection effects from professional football matches. To provide some evidence on the relevance of football matches for the infection development in a NUTS-3 district, we run an auxiliary regression which checks for the presence of mobility effects in hosting districts at match days. The idea is that football matches are only likely to have an observable impact at the population level if this event is sufficiently large to increase the general mobility level within a NUTS-3 district. Our auxiliary regression equation to test for mobility effects of professional football matches as a prerequisite for potential infection effects takes the following form

$$mobility_{i,t} = \gamma mday_{i,t} + \sum_{n=0} X_{i,t-n}\beta + week_t + dow_t + \mu_i + \upsilon_{i,t}, \tag{3.3}$$

where $mobility_{i,t}$ is a measure for mobility changes on a daily base as defined above and $mday_{i,t}$ is a nonabsorbing binary dummy, which is one if the NUTS-3 district hosts a professional football match on this day, i.e., $mday_{i,t} = 1[t = Match_i]$. Equation (3.3) also controls for the average temperature in the NUTS-3 district on day t and includes $t - n$ lagged values of mobility and the incidence rate (with $n = 3$). Finally $week_t$ and dow_t are week fixed effects and day-of-the-week fixed effects, respectively; μ_i are fixed effects for districts and $\upsilon_{i,t}$ is the model's error term.

4. EMPIRICAL RESULTS

4.1. Main findings

4.1.1. Incidence rates. Estimation results for the different DiD specifications presented in Table 2 generally point to statistically insignificant treatment effects from hosting professional football matches with live spectators on the regional COVID-19 infection dynamics (evaluated at a 10% critical level). This result is obtained by all three DiD specifications for the full sample including all professional football matches in our event window up to October 18, 2020.[15] However, when

[14] We conduct all SCM estimations in Stata using the SYNTH (Abadie et al., 2010) and SYNTH_RUNNER (Galiani and Quistorff, 2017) ado packages.

[15] Underlying probit regression results for DR estimation as well as balancing tests for covariates can be found in the Online Appendix (Table S2 and Figure S2).

we split the sample by leagues, the results in Table 2 point to positive, albeit only marginally significant, treatment effects for matches in the first league (Erste Bundesliga). In terms of effect size, the number of infections per 100,000 inhabitants is estimated to increase by about 0.5 to 0.6 cases per day in the first 14 days after the football match. Evaluated against the total number of infections in NUTS-3 regions with first league matches, this translates into an increase in daily incidence rates by about one twelfth (7.5 to 8%); effect size is very similar across the different specifications.[16]

Given that COVID-19 cases typically appear in the data with an average delay of roughly 7–10 days due to the duration until the onset of symptoms, testing procedures, and a reporting lag by health authorities (Mitze et al., 2020), we reevaluate the above findings by means of the PES approach. As shown in Figure 2, the plotted daily treatment effects largely confirm the DiD results. Panel A of Figure 2 visualizes the effects for all matches, showing that infections do not increase significantly after treatment start. Panel B for first league matches indicates that the daily incidence rate gradually increases for this subsample and that the treatment effect turns statistically significant in the second week after the match. For the overall sample, effect size is very similar if we compare the static DiD effect with the average of daily effects from PES estimation. In the case of the first league results, we find that the post-treatment average of daily PES effects are somewhat larger than the static baseline DiD effect. Given the potential bias of static DiD in the case of dynamically growing effects (Borusyak and Jaravel, 2020), we argue that the DiD estimates should be seen as conservative lower bounds for the true treatment effect.

The DiD and PES findings are further supported by SCM estimation. We find insignificant treatment effects for the full sample of match locations, but significant and sizeable effects for first league matches (see Figure 3). As for the dynamic PES approach, effects are observed to grow over time. Different from DiD and PES estimation, the SCM approach cannot explicitly control for a weekly pattern in the estimations (i.e., repeating spikes in reported COVID-19 cases over the different days during a calendar week as a result of the institutional setup of the German health system). We thus calculate daily marginal effects for SCM from a comparison of the seven-day rates, rather than the daily incidence rates, between treated districts and their respective synthetic control groups. While the temporal evolution of effects is thus somewhat smoother in the case of SCM compared to the PES results, the overall effect size and, importantly, dynamics is very similar across both estimation approaches.

4.1.2. Mobility estimates. While we found partial evidence for statistically significant treatment effects, particularly resulting from first league matches, the overall results for the full set of professional football matches were mainly insignificant. This raises the question whether effects are absent or whether the power of our district-level estimation strategy is simply too low to detect changes in epidemiological trends. To get an indication of the relevance of football matches and, thus, the estimation power of our approach, we run a series of auxiliary regressions to test for mobility effects as an alternative outcome (described at the end of Section 3). The idea is that if we do not detect changes in intra-district mobility on match days, professional football matches with spectators may be too small at the local population level to detect any district-wide effect. However, the regression results shown in Table 3 clearly point to statistically significant mobility effects on match days, which are both positive for the overall sample and the subsample of

[16] We also ran dynamic regressions using lagged incidence rates as explanatory variables—executed by the XTLSDVC (Bruno, 2005) package in Stata with an initial Blundell and Bond (1998) estimator. With regard to the estimated coefficients for the treatment indicator (shown in the Online Appendix Table S4), the coefficients remain quite similar in terms of effect size and statistical significance.

Table 2. DiD estimation results for COVID-19 infection effects of professional football matches with spectators in Germany.

Dep. Var.: $IR_{i,t}$	All matches (DFB cup and all leagues)				First league			
Specification	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)
$FTB_{i,t}$	0.280	0.244	0.230	0.286	0.595*	0.560*	0.575	0.656*
	(0.1885)	(0.1891)	(0.2424)	(0.2194)	(0.3115)	(0.3117)	(0.3714)	(0.3512)
Observations	9,003	18,747	18,747	18,747	8,023	17,767	17,767	17,767
Within R^2	0.32	0.24	0.31	0.31	0.33	0.24	0.32	0.32
Number of NUTS-3 districts	198	401	401	401	170	373	373	373
Region and day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend for region type I/II (linear and quadratic)	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Dep. Var.: $IR_{i,t}$	Second league				Third league			
Specification:	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)
$FTB_{i,t}$	0.011	−0.041	−0.003	0.084	0.253	0.218	0.201	0.285
	(0.3235)	(0.3238)	(0.3964)	(0.4254)	(0.3162)	(0.3187)	(0.3564)	(0.3434)
Observations	7,987	17,731	17,731	17,731	7,960	17,704	17,704	17,704
Within R^2	0.32	0.24	0.30	0.30	0.32	0.24	0.31	0.31
No. of NUTS-3 districts	169	372	372	372	169	372	372	372
Region and day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend for region type I/II (linear and quadratic)	No	Yes	Yes	Yes	No	Yes	Yes	Yes

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$; robust standard errors clustered at the regional level are given in parentheses; BS = bootstrap-based doubly robust DiD estimation with 250 bootstrap iterations.

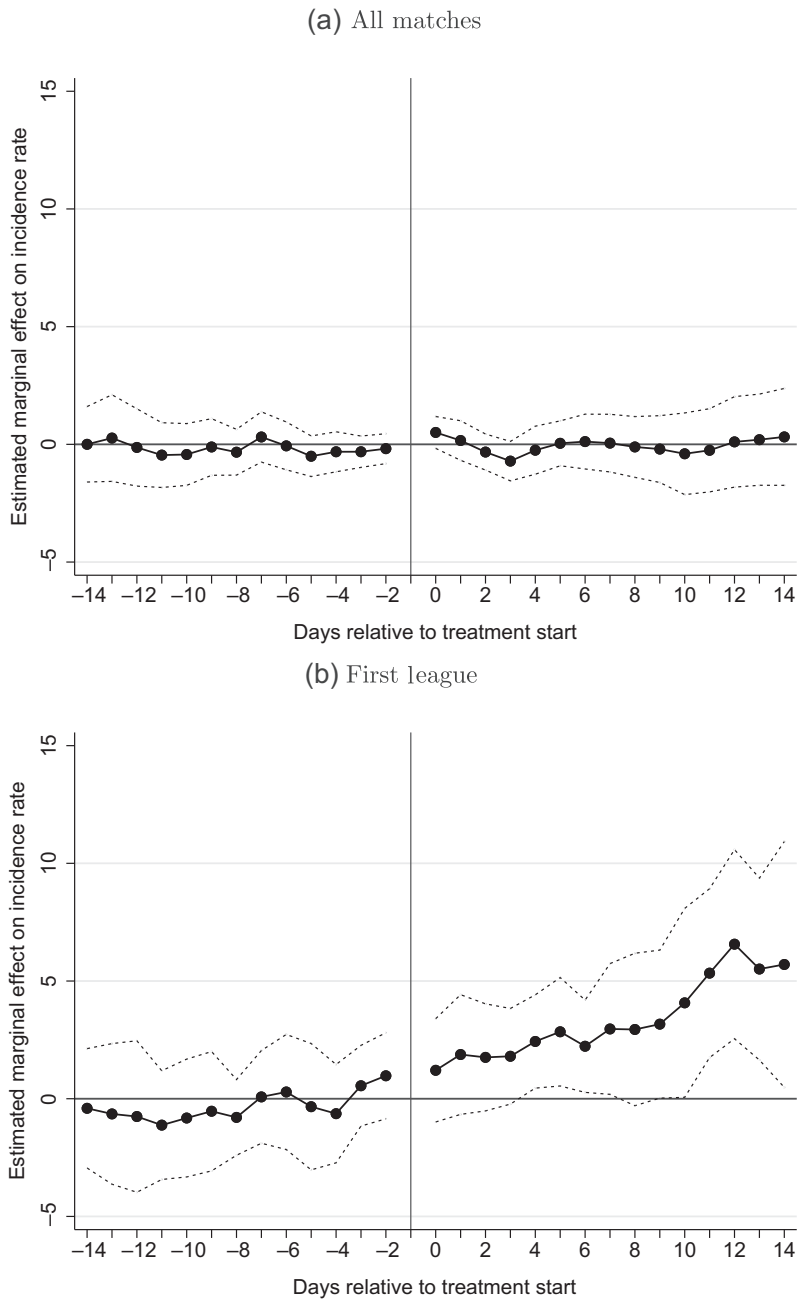

Figure 2. Estimated daily treatment effects from dynamic panel event study.
Notes: Black dots show point estimates, dashed lines indicate 99% confidence intervals. The last daily pretreatment dummy $FTB_{i,t}^{j}$ with $j = -1$ has been omitted to capture the baseline difference between treated and nontreated districts. Further details are given in the main text.

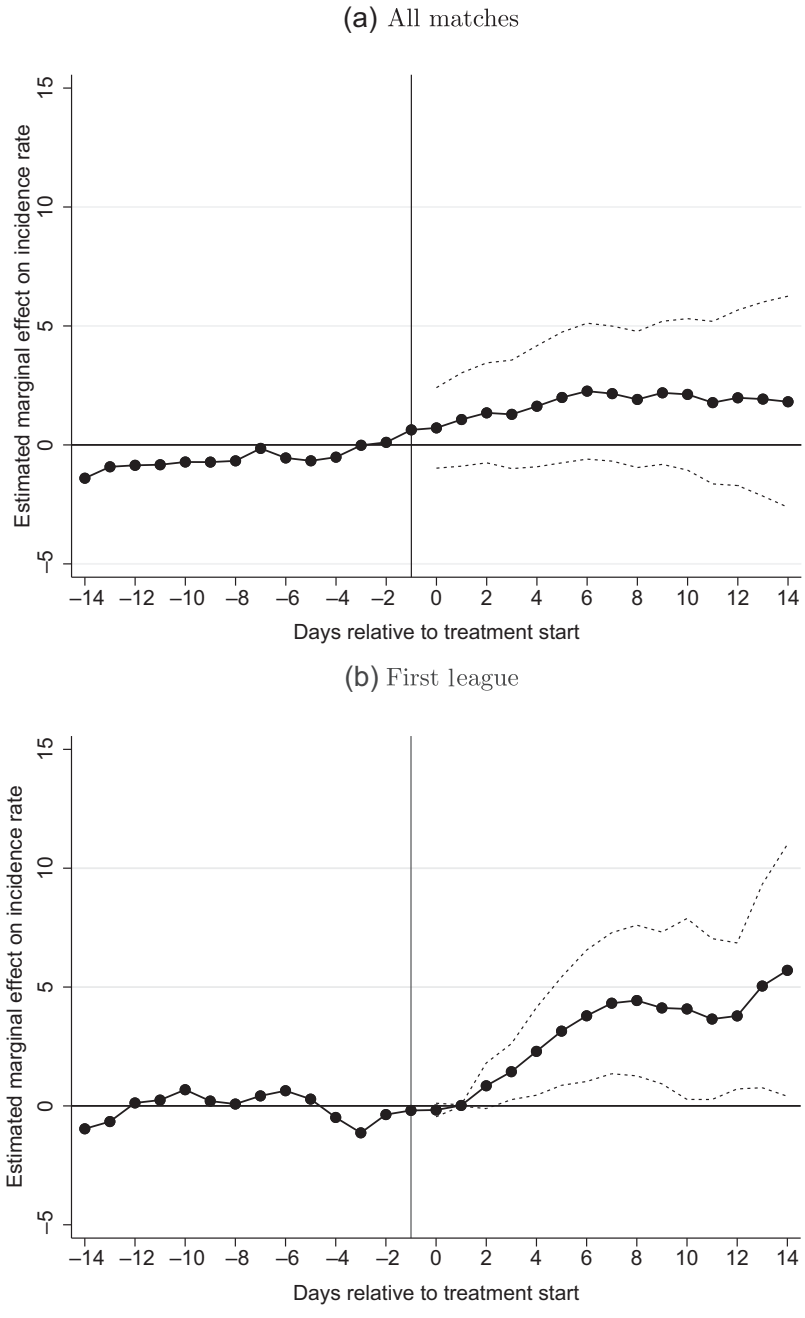

Figure 3. Estimated daily treatment effects from synthetic control method.
Notes: Black dots show the average difference in the seven-day incidence rate between treated districts and their respective synthetic control group; dashed lines indicate 95% confidence intervals calculated on the basis of match quality-adjusted pseudo *p*-values obtained from a series of placebo-in-space tests for the treatment period. Further details are given in the main text.

Table 3. DiD estimation results for mobility effects of professional football matches with spectators in Germany (overall and subsamples).

Dep. Var.: $mobility_{i,t}$	All matches			First league			First league		
Specification	Only urbanised	Trend for urbanised	Doubly robust	Only urbanised	Trend for urbanised	Doubly robust	Only urbanised	Trend for urbanised	Doubly robust
$mday_{i,t}$	0.020**	0.023**	0.014**	0.034**	0.038**	0.023**	0.029*	0.036**	0.020**
	(0.0092)	(0.0105)	(0.0060)	(0.0165)	(0.0190)	(0.0106)	(0.0164)	(0.0191)	(0.0100)
$mday_{i,t}$ (lead)	0.011	0.014	0.004	0.009	0.016	− 0.003			
	(0.092)	(0.0106)	(0.0053)	(0.0166)	(0.0190)	(0.0086)			
$mday_{i,t}$ (lag)	0.006	0.008	0.006	0.011	0.015	0.003			
	(0.0092)	(0.0105)	(0.0064)	(0.0165)	(0.0190)	(0.0090)			
$mday_{i,t}$ (away)							0.021	0.028	0.014
							(0.0177)	(0.0206)	(0.0112)
Observations	8,256	16,899	16,899	7,052	15,695	15,695	7,380	16,425	16,425
Within R^2	0.33	0.33	0.33	0.35	0.33	0.38	0.36	0.34	0.40
Number of NUTS-3 districts	393	393	393	164	365	365	164	365	365
Controls (daily temperature, lagged mobility, lagged incidence rate)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week and day-of-week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend for region type I/II (linear and quadratic)	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$; robust standard errors clustered at the regional level are given in parentheses. Bootstrap-based doubly robust DiD estimates are not reported here but can be obtained from the authors. Lead and lag terms cover the last two days before and after the match in treated NUTS-3 districts. The $mday_{i,t}$ (away) indicator measure mobility changes on match days in NUTS-3 districts being the home of the away team in a professional football match. The set of controls includes one- to three-day lags in mobility changes and incidence rates to capture (autoregressive) short-run dynamics.

first league matches. Importantly, if we conduct a series of placebo tests, e.g., by including a one- and two-day lead and lag into the model as well as testing for placebo effects in the origin districts of away teams for first league matches, we do not get evidence that our nonabsorbing match day dummy only captures a latent time trend. Taken together, the results for the mobility analysis in Table 3 do not underpin concerns on the estimation power of our population-level identification approach. They rather underline earlier findings that mobility changes may be an important channel for disease transmission (Bluhm and Pinkovskiy, 2021; Chernozhukov et al., 2020; Mitze and Kosfeld, 2021).

4.2. Robustness tests

This section serves to assess the robustness of our main findings on the infection effects of professional football matches. While the use of three alternative estimation methods can be seen as an important robustness test in itself, we additionally run a series of other tests to identify a potential sensitivity of the results. These additional tests differ by sample design related to the specification of the post-treatment period and the group composition of treated and nontreated districts. These robustness tests are mainly applied to the baseline DiD specification. As a first robustness test, we extend the treatment period from a maximum of 14 days in the basic DiD model to 20 days for all leagues (see columns 1–3 in Table 4) and for the first league separately (columns 4–6 in Table 4). The statistical significance of the estimated treatment effects is very similar for the extended treatment period as for the 14 days baseline estimates shown in Table 2. However, we observe that estimated effects are larger for the 20 days treatment period.

This finding is in line with the dynamic PES and SCM results pointing to accumulating effects over time. Nonetheless, we treat the results from this longer 20-days treatment period with some caution for the following reason. As most districts in the treatment group host a new match after two weeks, effect identification may suffer from the problem of multiple treatments for each treated unit. Moreover, extending the post-treatment period also increases the risk of capturing latent events as drivers of a region's infection dynamics that are not related to professional football matches in the coefficient for the treatment indicator. In our (robustness) analysis, we will thus primarily focus on the event window of 14 days after each individual football match.

We also run placebo regressions, assigning treatment status not to each de facto treated NUTS-3 district, but to the district of the away team in each match, which is not 'treated' as the aforementioned hygiene protocols did not allow away fans in the stadium. Any latent regional factors (specific characteristics that only apply for districts with matches) that confound our identification strategy, should lead to the same results in the placebo regressions. Or, put differently, significant treatment, but insignificant placebo treatment, effect would tell a clear story in terms of the infection effects from professional football matches with spectators. Columns 7–9 in Table 4 show the results from these placebo regressions that estimate pseudo-treatment effects for the subsample of first league matches, for which we have identified significant treatment effects. The placebo treatment results reported in Table 4 are statistically insignificant across all regression specifications. Hence, the main result from this placebo test is that we do not find evidence for latent confounding factors, which correlate with the status of hosting a first league match, biasing our estimation results.

This finding further places the focus on potential infections in the stadium. One common argument made in the public debate is that—given the large fan base of first league teams—not stadium visits as such are a risk for higher infection rates, but rather private meetings of fans to watch the match on television, either in private locations, restaurants, or bars. The results from

Table 4. Robustness tests for treatment period for 20 days and placebo treatment effects for NUTS-3 districts of away teams.

Dep. Var.: $IR_{i,t}$	Treatment period: 20 days								Placebo treatment[a]			
	All matches				First league				First league			
Specification	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)
$FTB_{i,t}$	0.372	0.362	0.217	0.306	0.890**	0.885**	0.704	0.808*				
	(0.2293)	(0.2285)	(0.2982)	(0.2809)	(0.3764)	(0.3753)	(0.4600)	(0.4423)				
Placebo $FTB_{i,t}$ (away)									0.284	0.249	0.289	0.315
									(0.2983)	(0.2962)	(0.4170)	(0.3590)
Observations	9,249	18,993	18,993	18,993	8,101	17,845	17,845	17,845	7,697	17,441	17,441	17,441
Within R^2	0.34	0.25	0.37	0.37	0.33	0.25	0.36	0.36	0.33	0.24	0.32	0.32
No. of NUTS-3 districts	198	401	401	401	170	373	373	373	161	364	364	364
Region and day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time trend for region type I/II (linear and quadratic)	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$ robust standard errors clustered at the regional level are given in parentheses. BS = bootstrap-based doubly robust DiD estimation with 250 bootstrap iterations. For the placebo treatment regressions, the first step probit model of the doubly robust DiD specification has been the estimator on the basis of a binary dummy that takes a value of one for NUTS-3 districts hosting the away team of a professional football match on the first match day of the first league.
[a] Berlin is excluded in the regressions because Berlin has two professional football clubs in the first league who had their home match on different match days. We cannot assign a placebo treatment to Berlin on any match day as it is treated on both match days.

our placebo analysis do not provide evidence for this transmission channel, though, as we should observe similar effects in treated districts hosting the match and in the home district of the away team. This is not the case.

We have also tested for potential spillover effects of football matches to neighbouring districts, i.e., districts adjacent to districts hosting a match. Our identification strategy so far has been built on the assumption that spectators are solely inhabitants of the NUTS-3 districts hosting the match (i.e., COVID-19 transmission are restricted to take place within this district). As an extension, we have also tested whether the incidence rates in neighbouring districts of match locations are similarly affected by the matches or not (results are reported in columns 1–2 in Table S5 in the Online Appendix). Additionally, we have tested if omitting these neighbours affects the baseline DiD results. If potential latent spillover effects to neighbouring districts affect the control group, such spillovers would bias the estimated effects downwards (results are reported in columns 3–4 in Table S5 in the Online Appendix). Taken together, we find that treatment effects in neighbouring regions of a match location do not appear to be statistically significant. Deleting neighbouring counties also does not affect the baseline estimation results.

Two further aspects are considered for additional robustness tests. First, we focus on the group of treated districts with first league matches for which we find significant treatment effects. As this group is relatively small, observed effects may thus be driven by the infection development in one specific 'hot spot'. Although Figure S4 in the Online Appendix shows that all regions with a professional football match lie within the overall regional distribution of the seven-day incidence rate, the corresponding rates for NUTS-3 districts hosting a first league match tend to vary stronger over time compared to the epidemiological development of match locations for the second and third league. To rule out that our estimated effects are driven by single districts, we perform leave-one-out estimates. In doing so, we run 14 regressions, sequentially leaving out one treated first league match district and test whether the results turn out insignificant. Although the effect size varies between the different regressions, the statistical significance remains mostly stable throughout this leave-one-out analysis.[17]

Second, we leave out those districts that have exceeded a seven-day incidence rate of 35 infections per 100,000 inhabitants during our sample period. The reason is that this serves as a first critical benchmark to initiate specific local measures by public health authorities to suppress the further spreading of COVID-19 in these regions. Including such districts in the comparison group may potentially bias the average infection rate in this group upwards (if dynamic effects are present) or downwards (if initiated measures of local authorities additionally slowdown incidence rates in these regions). Turning to the results of this last robustness test (see Table S3 in the Online Appendix for detailed regression outputs), we find smaller effects compared to the benchmark treatment effects (as shown in Table S6 in the Online Appendix) together with mixed evidence for statistical significance.

4.3. Transmission channels

Going beyond the overall identification of treatment effects, we also study underlying channels offering potential causes for effects to occur. The empirical evidence for statistically significant treatment effects in the first league, but insignificant findings for other matches, poses the question of the underlying transmission channels that may drive this result. Basically, we apply heterogeneity analyses, interacting the treatment indicator with further indicators which may plausibly

[17] Results are not reported here, but are available on request and can be run using the replication files accompanying this paper.

cause higher infections after matches. Equation (4.1) shows the underlying test strategy for the identification of transmission channels in the baseline DiD model.

For each of the different channels we test for differences in the coefficients δ_1 and δ_2 for the interaction terms $\Psi_{i,low} \times FTB_{i,t}$ and $\Psi_{i,high} \times FTB_{i,t}$. Thereby $\Psi_{i,low}$ is an indicator for matches with below median for (a) number of spectators, (b) utilisation rates, (c) probability of using public transport, and (d) the strictness of face mask regulations, respectively (we do not expect strong effects for matches indicated with $\Psi_{i,low}$). Vice versa, $\Psi_{i,high}$ is an indicator for matches where we expect significant effects related to above median values for one of the abovementioned transmission channels.[18]

$$IR_{i,t} = \delta_1(\Psi_{i,low} \times FTB_{i,t}) + \delta_2(\Psi_{i,high} \times FTB_{i,t}) + X_{i,t}\beta + \tau_t + \mu_i + \varepsilon_{i,t}. \tag{4.1}$$

To be more specific, the selection of potential transmission channels builds on the following ideas: first, we test if the total number of spectators affects the results. This approach builds on the hypothesis that, especially around large stadiums, facilities are crowded if more spectators are permitted to watch the match. Focusing on first league matches, we construct a binary dummy that takes a value of 1 for those matches in which the number of spectators exceeds the sample median (approx. 6,300 spectators). We label this dummy 'large match'. We then interact this dummy with the basic treatment dummy and do the same for matches with the number of spectators below median ('small match'). The coefficient for treated regions with an above median number of spectators is found to be statistically significant and remains stable over different DiD specifications with an effect size of about 0.8 to 0.9 (Table 5, columns 1–3). However, if we test for coefficient equality between both treatment groups, the null hypothesis of equal coefficients cannot be rejected. This points to weak differences between matches with an above and below median number of spectators.

Second, we seek to detect if utilisation rates within stadiums (and not the total number of spectators) affect the risk of higher infections. Utilisation rates are measured as such that they relate the number of spectators to total stadium capacity (seat occupation). This approach follows the idea that stadiums with a relatively high seat occupation rate may be too crowded to ensure adequate social distancing when entering/exiting and moving around in the stadium. As before, we capture potentially heterogeneities in the treatment group by forming separate dummies for matches with a spectator density above/below the median. However, the estimates generally do not show a significantly higher incidence rate for one of these two subgroups (see Table 5) and coefficient equality across groups cannot be rejected.

Third, in a similar vein, we test whether the share of spectators that access the stadium by public transport increases the infection rate. As we do not have any data on actual public transport use, we proxy the probability of using public transport by the car parking capability at each stadium. This proxy builds on the hypothesis that a lower probability to travel by car (in nonpandemic times) is indicated by fewer parking capabilities and that this link also holds for travel behaviour during the COVID-19 pandemic. However, our estimates shown in Table 5 do not provide evidence for this transmission channel. The results remain statistically insignificant throughout all specifications.

Lastly, we exploit differences in hygiene protocols in each stadium with regard to specific face mask regulations. We can identify one distinct difference in these protocols. One type of protocol mandates the wearing of face mask when entering or walking around in the stadium, but does not require wearing a mask when seated (we refer to this as 'limited face mask obligation'). The other, stricter protocol mandates the wearing of face masks throughout the entire stadium visit

[18] An overview on the relevant indicators used for the robustness analysis is given in Table S3 in the Online Appendix.

Table 5. Identification of COVID-19 transmission channels based on differences in stadium characteristics (first league matches).

Dep. Var.: $IR_{i,t}$	Number of spectators				Seat occupation rate			
Specification	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)
$FTB_{i,t} \times small\ match$	0.369	0.348	0.275	0.339				
	(0.4090)	(0.4120)	(0.4628)	(0.4474)				
$FTB_{i,t} \times large\ match$	0.892**	0.839*	0.959**	1.092**				
	(0.4330)	(0.4305)	(0.4824)	(0.5234)				
$FTB_{i,t} \times low\ occupation$					0.683	0.654	0.667	0.641
					(0.5876)	(0.6034)	(0.5790)	(0.6314)
$FTB_{i,t} \times high\ occupation$					0.579*	0.543	0.559	0.567
					(0.3455)	(0.3451)	(0.3982)	(0.3896)
Observations	8,023	17,767	17,767	17,767	8,023	17,767	17,767	17,767
Wald F-test for	F(1,169)	F(1,372)	F(1,372)	–	F(1,169)	F(1,372)	F(1,372)	–
coefficient equality	0.81	0.71	1.27		0.88	0.87	0.86	
Within R^2	0.33	0.24	0.32	0.32	0.33	0.24	0.32	0.32
No. of NUTS-3 districts	170	373	373	373	170	373	373	373
Dep. Var.: $IR_{i,t}$	Public transport				Face mask obligation			
Specification:	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)	Only urbanised	Trend for urbanised	Doubly robust	Doubly robust (BS)
$FTB_{i,t} \times low\ publ.\ transport$	0.484	0.452	0.458	0.526				
	(0.3974)	(0.3946)	(0.4497)	(0.5012)				
$FTB_{i,t} \times high\ publ.\ transport$	0.723	0.683	0.709	0.799*				
	(0.4636)	(0.4674)	(0.5050)	(0.4784)				
$FTB_{i,t} \times strict\ face\ mask$					−0.101	−0.149	−0.110	−0.038
					(0.1929)	(0.2021)	(0.2946)	(0.2276)
$FTB_{i,t} \times limited\ face\ mask$					0.899**	0.869**	0.877**	0.955**
					(0.3936)	(0.3912)	(0.4416)	(0.4721)
Observations	8,023	17,767	17,767	17,767	8,023	17,767	17,767	17,767
Wald F-test for	F(1,169)	F(1,372)	F(1,372)	–	F(1,169)	F(1,372)	F(1,372)	–
coefficient equality	0.16	0.15	0.17		5.78**	5.91**	5.65**	
Within R^2	0.33	0.24	0.32	0.32	0.33	0.24	0.32	0.32
Number of NUTS-3 districts	170	373	373	373	170	373	373	373

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$; robust standard errors clustered at the regional level are given in parentheses. BS = bootstrap-based doubly robust DiD estimation with 250 bootstrap iterations. Region and day FE are included in all cases. Time trends (linear and quadratic) for region type I/II are included in the estimations except for the 'Only urbanised' subsample.

without any exception. Prior research has pointed to the role of face masks suppressing the spread of the virus (Chernozhukov et al., 2020; Mitze et al., 2020). Accordingly, we construct a dummy that takes a value of 1 for matches with limited mask obligation, i.e., prescribe the wearing of masks when walking around in the stadium but not when seated.[19] This dummy is also interacted with the treatment dummy.

The result for the interaction term between the limited face masks dummy and the treatment indicator in this case shows that NUTS-3 districts hosting a match only with a limited face mask obligation experience a significantly higher COVID-19 incidence rate in the treatment period of 14 days following the match. In terms of effect size, the increase translates into a higher infection rate of 0.8 to 1.0 cases per 100,000 inhabitants per day. This translates into an approximately 20% relative rise in the incidence rate for treated regions during the treatment period. For districts hosting a match with strict face masks regulations no increases in the incidence rate are observed.

The face masks regulations mark the most important transmission channel identified from the interaction term estimates (in terms of statistical significance and effect size). In order to further investigate potential dynamic effects, we also apply PES and SCM to this specific channel. The obtained results are shown in Panel A and Panel B of Figure 4. In the case of PES estimation, we find that football matches with limited face mask obligations clearly show that marginal effects on the incidence rate become statistically significant after approximately 10 days and are larger than the overall results for first league matches. However, interpretations for this specific result should be done carefully because the visual inspection of the figure indicates that the common trend assumption may not hold consistently, i.e., estimated effects prior to treatment start seem to pick up a trend that continues during the treatment period. Still, we observe that, only about nine days after treatment start, incidence rates in treated districts significantly exceed those in nontreated comparison districts (evaluated against the 95% confidence intervals plotted in the figure).

The SCM estimation (shown in Panel B of Figure 4) confirm statistically significant infection effects and do not indicate any latent trend starting in the pretreatment period. Additional PES estimates for matches with an above median number of spectators are shown in the Online Appendix (Figure S3, Panel B). Here, the PES results confirm the DiD estimates that treatment effects are only marginally significant over the 14-day treatment window. Taken together, our estimates for the identification of potential transmission channels of infection effects from stadium visits identify a strict face mask requirement as an important element of public health strategies and associated hygiene protocols in times of the COVID-19 pandemic with no little alternative means to control for latent infection risks.

5. CONCLUSION AND DISCUSSION

With the start of the 2020/2021 season, German professional football teams, supported by German politics and health authorities, initiated an 'experimental' phase that reopened stadiums for up to 10,000 spectators per match. Reopening was possible if certain epidemiological conditions were met and organising clubs implemented hygiene protocols approved by local health authorities. In this context, we have investigated the effects of professional football matches with at least 1,000 live spectators on population-level COVID-19 incidence rates in NUTS-3 districts hosting

[19] When we compare the seven-day incidence rates between hosting districts with and without strict face mask obligations, this does not show significant mean differences on the basis of *t*-tests. Hence, stricter face mask obligations do not seem to be a reflex of higher incidence rates and accordingly stricter local policy measures.

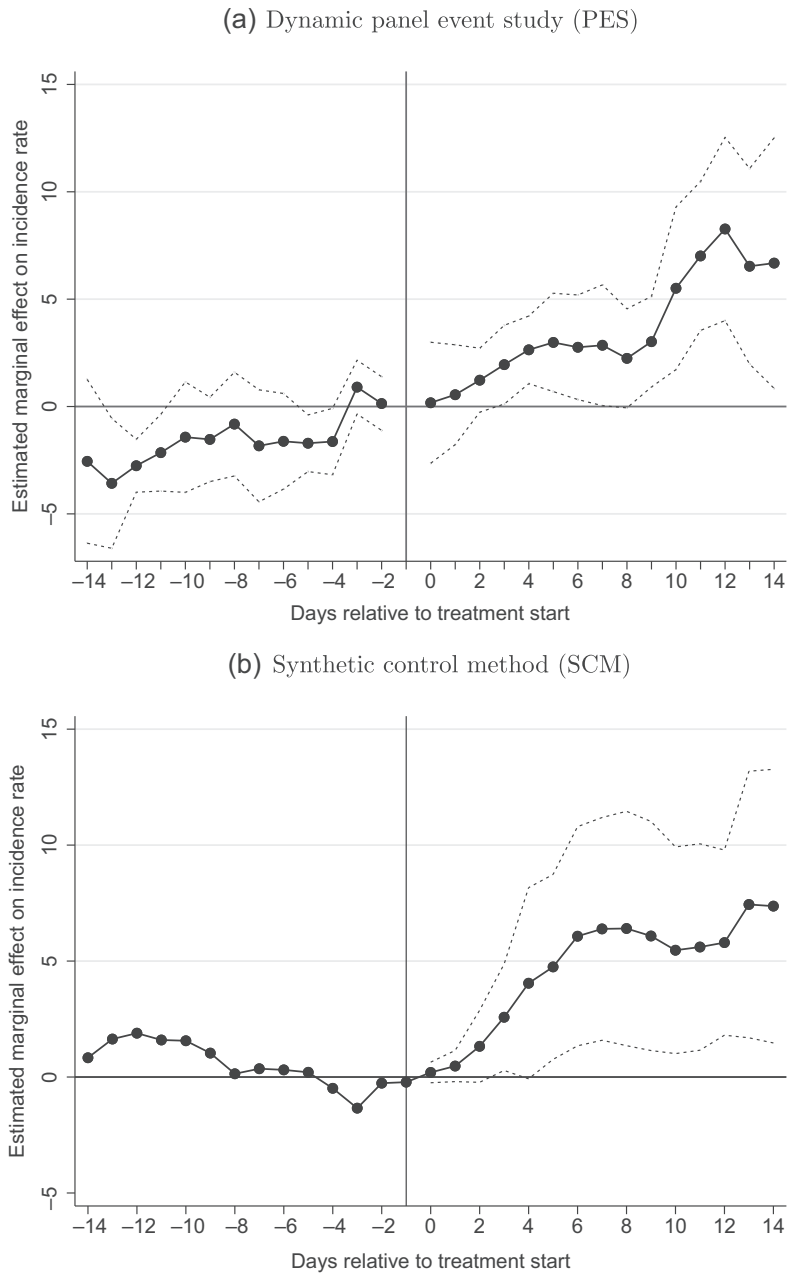

Figure 4. Treatment effects for first league matches with limited face mask requirement.
Notes: Black dots show estimated treatment effect per day, dashed lines indicate 95% confidence intervals. In the case of the event study, the last pretreatment dummy $FTB^{j}_{i,t}$ with $j = -1$ has been omitted to capture the baseline difference between treated and nontreated districts. In the case of SCM estimation, confidence intervals are calculated on the basis of match quality-adjusted pseudo p-values obtained from a series of placebo-in-space tests for the treatment period. Further details are given in the main text.

a match using a quasi-experimental setup. Our sample with daily observations for 401 NUTS-3 regions covers the period from August 10 to October 18, 2020, with matches taking place throughout September in the first round of the national cup and the first two match days of all three professional football leagues. This clearly defined event window allows us to observe potential infection effects for at least 14 days after the match.

We have estimated treatment effects by static difference-in-difference estimation with staggered treatment start, dynamic panel event study, and synthetic control method. For the full sample of 41 match locations, our estimation results do not provide evidence for statistically significant higher infection dynamics in NUTS-3 districts hosting a match vis-à-vis similar district without this treatment. However, we find evidence for significant treatment effects once we limit the sample to first league matches. Our findings point to a significant increase in the infection rate roughly 1–2 weeks after the match. To assess the power of our estimation approach, we have also tested for mobility effects in match locations on match days. We find a significant increase in mobility not only for first league matches but also for the overall sample. This makes us confident that the obtained results at the local population level do not suffer from a low estimation power and that effects may differ by professional football leagues.

To investigate potential transmission channels we have, among other factors, checked for treatment heterogeneities stemming from match size, i.e., the number of spectators, and from different hygiene protocols. Matches with more spectators seem to be associated with a higher incidence rate. However, given that post-estimation tests cannot reject coefficient equality between larger and smaller matches, results should be interpreted with care and we argue that this transmission channel is weak. Moreover, our results from all three estimation methods clearly suggest that a rigorous face mask obligation should be implemented for future matches with live spectators. Here, results point to a significant regional COVID-19 increase if hygiene protocols only provided limited face mask covering (i.e., wearing masks while walking around inside the stadium but not when seated). This result is not observed for matches with a strict face mask obligation throughout the entire stadium visit.[20]

Taken together, although no systematic effects for all matches could thus be observed, we conclude that certain large-scale sport events, i.e., particularly matches with a large fan base that accordingly mobilise a large number of people to watch these events live, may pose an additional COVID-19 infection risk under the general circumstances studied. This effect becomes especially visible when estimators are applied that account for the dynamic, i.e., time heterogeneous, development of incidence rates at the local population level.

We have carefully tried to evaluate the sensitivity of our findings to changes in the estimation setup and sample design. We find that the results are robust against the chosen estimation method and also the conduct of placebo regressions, which test for effects in the home districts of away teams in individual matches (while fans from these teams were not allowed to enter the stadium). The placebo regressions thus exploit the random distribution of match days between the home and away teams. In this setting, it can be assumed that the treatment group in the placebo regressions hardly differs from the group of districts hosting a match with spectators on the respective match days. Strengthening our finding, these results from the placebo test remain statistically insignificant without exception (in contrast to the actual treatment effects) and do not have positive or significant coefficients.

[20] Some clubs have already anticipated this potential transmission channel and changed their hygiene protocol (after our sample period has ended). See a press report by the German-wide sports news portal *Kicker* at: https://www.kicker.de/vfl-appell-vor-bielefeld-maske-auf-788051/artikel (last accessed: November 19, 2020).

Our district-level identification approach has the natural limitation that we cannot conclude that infections took place right away in the stadium. For example, the additional infection effects observed during match days with many spectators may indicate that bottlenecks are created in the vicinity of the stadiums, leading to additional infections. On a more general level, football matches taking place with many live spectators may also indirectly affect the compliance with hygiene rules and social distancing (e.g., private meetings) of other inhabitants. People may be less willing to restrict their own behaviour when they see spectators joining a large event on their way to the stadium. These factors may also explain higher incidence rates in match locations.

Based on our findings, the best advice that we can give to public health authorities and organisers of future large-scale sport events in times of the COVID-19 pandemic is that proper hygiene protocols should be a key element of the reopening concepts. Specifically, assigned seating, social distancing measures in the stadium and its vicinity and, most importantly on the basis of our findings, an extensive face mask duty should be applied also in times of lower incidence rates (see Tupper et al., 2020, for similar suggestions in the context of comparable events). These measures should be accompanied by systematic use of rapid testing for stadium visitors—an extension to existing hygiene protocols, which was not available during the 'experimental' phase in late summer and early autumn 2020. If combined with increasing vaccination rates, this should open up the possibility to let fans back into stadiums in an opening-under-safety approach.

In addition (not testable in our data), fans and spectators, in- and outside the stadium, need to understand that compliance to hygiene protocols is essential for limiting the spread of the virus. We also need to state that our observed results have to be interpreted conditional to the overall national infection dynamics, which was moderate in Germany for the sample period under investigation in this study, i.e., our nonsignificant results (except for first league matches) should not be taken at face value as evidence for reopening stadiums during the peak of a pandemic wave and/or the emergence of virus mutations that may be associated with a significantly higher viral transmissibility even for outdoor activities.

DATA AND CODE AVAILABILITY

All study data have been made public to support replication studies. Data files and Stata codes are provided as supplementary data and can also be accessed in the following data repository with DOI: 10.6084/m9.figshare.13337966.

ACKNOWLEDGEMENTS

We thank the two anonymous reviewers for their valuable comments throughout the publication process. We also thank Thomas Bauer, Sandra Schaffner, and Klaus Wälde for their valuable advice and comments on earlier versions of this paper. We are grateful to Thorben Wiebe for very helpful research assistance.

CONFLICT OF INTEREST STATEMENT

Both authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

REFERENCES

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*(1), 1–19.

Abadie, A. (2020). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature 59*, 391–425.

Abadie, A., A. Diamond and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association 105*(490), 493–505.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review 93*, 113–32.

Altman, D. G. and J. M. Bland (2011). How to obtain the confidence interval from a p-value. *BMJ 343*, d2090.

Athey, S. and G. Imbens (2018). Design-based analysis in difference-in-differences settings with staggered adoption, Working paper 24963, National Bureau of Economic Research, Cambridge, MA.

BBSR (2021). INKAR Indikatoren und Karten zur Raum- und Stadtentwicklung. Federal Institute for Research on Building, Urban Affairs and Spatial Development, Available at: https://www.inkar.de (last accessed: June 24, 2021).

Bluhm, R. and M. Pinkovskiy (2021). The spread of COVID-19 and the BCG vaccine: A natural experiment in reunified Germany. *Econometrics Journal*, Published ahead of print, May 7. https://doi.org/10.1093/ectj/utab006.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–43.

Bodory, H., L. Camponovo, M. Huber and M. Lechner (2020). The finite sample performance of inference methods for propensity score matching and weighting estimators. *Journal of Business and Economic Statistics 38*(1), 183–200.

Borusyak, K. and X. Jaravel (2020). Revisiting event study designs. Working paper, Harvard University, Cambridge, MA. Available at: https://scholar.harvard.edu/borusyak/publications/revisiting-event-study-designs (April 25, 2020).

Bruno, G. (2005). XTLSDVC: Stata module to estimate bias corrected LSDV dynamic panel data models. Statistical software components s450101, Boston College Department of Economics, Boston, MA.

Cavallo, E., S. Galiani, I. Noy and J. Pantano (2013). Catastrophic natural disasters and economic growth. *Review of Economics and Statistics 95*, 1549–61.

Chernozhukov, V., H. Kasahara and P. Schrimpf (2020). Causal impact of masks, policies, behavior on early COVID-19 pandemic in the US. *Journal of Econometrics 220*(1), 23–62.

Cho, S. W. (2020). Quantifying the impact of nonpharmaceutical interventions during the COVID-19 outbreak: The case of Sweden. *Econometrics Journal 23*(3), 323–44.

CoronaSchVO(2020). Verordnung zum Schutz vor Neuinfizierungen mit dem Coronavirus SARS-CoV-2, July 1, 2020. Ministry of Labour, Health and Social Affairs for North Rhine-Westphalia, Düsseldorf, Germany. Available at: https://www.land.nrw/sites/default/files/asset/document/2020-07-01_coronaschvo_vom_01.07.2020.pdf (last accessed: March 31, 2021).

Dave, D. M., A. I. Friedson, K. Matsuzawa, J. J. Sabia and S. Safford (2020). Black lives matter protests, social distancing, and COVID-19. Working paper 27408, National Bureau of Economic Research, Cambridge, MA.

Davies, N. G., S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, et al (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science 72*(6538), eabg3055.

De Bruin, Y. B., A. S. Lequarre, J. McCourt, P. Clevestig, F. Pigazzani, M. Z. Jeddi and M. Goulart (2020). Initial impacts of global risk mitigation measures taken during the combatting of the COVID-19 pandemic. *Safety Science 128*, 104773.

Destatis (2021). Experimental data: Mobility indicators based on mobile network data. Statistisches Bundesamt, Wiesbaden, Germany. Available at: https://www.destatis.de/EN/Service/EXDAT/Datensaetze/mobility-indicators-mobilephone.html (last accessed: June 24, 2021).

DFB (2021a). Ligen und wettbewerbe. Available at: https://www.dfb.de/ligen-wettbewerbe (last accessed: June 24, 2021).

DFB (2021b). Vereins-Finder, Datencenter. Available at: https://www.dfb.de/datencenter/vereine (last accessed: June 24, 2021).

Drewes, M., F. Daumann and F. Follert (2021). Exploring the sports economic impact of COVID-19 on professional soccer. *Soccer and Society 22*(1–2), 125–37.

DWD (2021). Climate data center. Available at: https://www.dwd.de/EN/climate_environment/cdc/cdc_node_en.html (last accessed: June 24, 2021).

Freedman, D. A. and R. A. Berk (2008). Weighting regressions by propensity scores. *Evaluation Review 32*(4), 392–409.

Friedson, A. I., D. McNichols, J. J. Sabia and D. Dave (2021). Shelter-in-place orders and public health: Evidence from California during the COVID-19 pandemic. *Journal of Policy Analysis and Management 40*, 258–83.

Funk, M. J., D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart and M. Davidian (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology 173*(7), 761–7.

Galiani, S. and B. Quistorff (2017). The synth_runner package: Utilities to automate synthetic control estimation using synth. *Stata Journal 17*, 834–49.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Working paper 25018, National Bureau of Economic Research, Cambridge, MA.

Goodman-Bacon, A. and J. Marcus (2020). Using difference-in-differences to identify causal effects of COVID-19 policies. *Survey Research Methods*, 153–58.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998). Characterizing selection bias using experimental data. *Econometrica 66*, 1017–98.

Isphording, I., M. Lipfert and N. Pestel (2020). School re-openings after summer breaks in Germany did not increase SARS-CoV-2 cases. Discussion paper 13790, IZA Institute of Labor Economics, Bonn, Germany.

Lechner, M. (2010). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics 4*(3), 165–224.

Mangrum, D. and P. Niekamp (2020). JUE insight: College student travel contributed to local COVID-19 spread. *Journal of Urban Economics*, Published ahead of print, December 4. https://doi.org/10.1016/j.jue.2020.103311.

Meyer, T., D. Mack, K. Donde, O. Harzer, W. Krutsch, A. Rössler, J. Kimpel, D. von Laer and B. C. Gärtner (2021). Successful return to professional men's football (soccer) competition after the COVID-shutdown: A cohort study in the German Bundesliga. *British Journal of Sports Medicine 19*(55), 62–66.

Mitze, T. and R. Kosfeld (2021). The propagation effect of commuting to work in the spatial transmission of COVID-19. *Journal of Geographical Systems*, Published ahead of print, May 23. https://doi.org/10.1007/s10109-021-00349-3.

Mitze, T., R. Kosfeld, J. Rode and K. Wälde (2020). Face masks considerably reduce COVID-19 cases in Germany. *Proceedings of the National Academy of Sciences 117*(51), 32293–301.

Mitze, T. and J. Rode (2021). Early assessment of epidemiological trends associated with SARS-CoV-2 variants of concern in Germany. Working paper, medRxiv. https://doi.org/10.1101/2021.02.16.21251803.

Parnell, D., P. Widdop, A. Bond and R. Wilson (2020). COVID-19, networks and sport. *Managing Sport and Leisure*, Published ahead of print, March 31. https://doi.org/10.1080/23750472.2020.1750100.

RKI (2020a). Infektionsumfeld von COVID-19-Ausbrüchen in Deutschland. Epidemiologisches Bulletin 38, Robert Koch Institute, Berlin.

RKI (2020b). Täglicher lagebericht des RKI zur coronavirus-krankheit-2019 (COVID-19). Robert Koch Institute, Berlin.

RKI (2021). COVID-19 Datenhub. Available at: https://npgeo-corona-npgeo-de.hub.arcgis.com (last accessed: June 24, 2021).

Sant'Anna, P. H. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics 219*(1), 101–22.

Schlosser, F., B. F. Maier, O. Jack, D. Hinrichs, A. Zachariae and D. Brockmann (2020). Covid-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proceedings of the National Academy of Sciences 117*(52), 32883–90.

Schmidheiny, K. and S. Siegloch (2019). On event study designs and distributed-lag models: Equivalence, generalization and practical implications. Working Paper 7481, CESifo, Munich, Germany.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, Published ahead of print, December 16. https://doi.org/10.1016/j.jeconom.2020.09.006.

Tupper, P., H. Boury, M. Yerlanov and C. Cloijn (2020). Event-specific interventions to minimize COVID-19 transmission. *Proceedings of the National Academy of Sciences 117*, 32038–45.

Von Bismarck-Osten, C., K. Borusyak and U. Schönberg (2020). The role of schools in transmission of the SARS-CoV-2 Virus: Quasi-experimental evidence from Germany. Discussion paper 2022, Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London, UK.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Victor Chernozhukov handled this manuscript.